Two-dimensional Contingency Tables with Both Completely and Partially Cross-classified Data

by

Tar Chen* and Stephen E. Fienberg#

Technical Report No. 211

May 1973

*Illinois Bell Telephone Co., Chicago, Illinois.

#Department of Applied Statistics, University of Minnesota, St. Paul, Minnesota.

Summary

Models are developed for the analysis of contingency table data with supplemental marginal totals. The method of maximum likelihood is used to estimate the parameters in the models, and the expected cell values for goodness-of-fit statistics. The value of utilizing the supplemental margins is discussed in terms of asymptotic variances and the consistency of estimates. The approach developed is illustrated in a 2 x 2 table example.

1. Introduction

In the analysis of contingency tables it frequently happens, either by chance or by design, that some observations are only partially cross-classified, according to either the row classification or the column classification. Blumenthal [1968] and Hocking and Oxspring [1971] have considered estimating multinomial cell probabilities using maximum likelihood in cases where some of the data are partially classified, but they do not deal with the special structure of cross-classified data, nor reduced parametrizations of general interest, e.g. independence of the variables corresponding to rows and columns. These authors view the partially classified data as consisting of multinomial observations with parameters related to those of the parameters of the multinomial distributed completely classified data. Such an approach is a special case of the one we adopt in this paper. Reinfurt [1970], Koch and Reinfurt [1970], and Koch, Imrey and Reinfurt [1972] use a modified minimum chi-squared approach to the contingency table version of the Blumenthal-Hocking-Oxspring problem.

In this paper we describe methods for obtaining maximum likelihood estimates for expected cell values in contingency tables with partially cross-classified data. We consider two models for the basic cross-classification (unrestricted and independence) and a special class of "random" mechanisms which produce the partially cross-classified data, explaining how they relate to the completely cross-classified data. First we obtain maximum likelihood estimates for the parameters of the different models, then we use these to get estimated expected cell values and associated goodness-of-fit statistics. We also discuss the use of partitioning of these statistics in order to select as simple a model as possible to describe the observed data.

In a subsequent paper we will extend the methods considered here to deal with estimation in multi-dimensional contingency tables with some partially cross-classified data, using loglinear models (see Birch [1963] and Bishop [1969]).

2. Notation

Let x_{ij} denote the fully cross-classified count for the (i,j) cell of an r x c contingency table, R_i $(i = 1,2,\ldots,r)$ the count of the partially classified individuals corresponding to the i-th row, and C_j $(j = 1,2,\ldots,c)$ the count of the partially classified individuals corresponding to the j-th column. We refer to the $\{R_i\}$ and $\{C_j\}$ as supplemental margins or partially cross-classified counts (see Koch, Imrey and Reinfurt [1972]). The total sample size is

$$\begin{aligned} N &= \sum_{ij} x_{ij} + \sum_i R_i + \sum_j C_j \\ &= x_{++} + R_+ + C_+ \ . \end{aligned} \tag{1}$$

We assume that $x_{ij} > 0$ for $i = 1,2,\ldots,r$ and $j = 1,2,\ldots,c$.

We envision a two-stage procedure for allocating observations to their observed cells. In the first stage, which is never observed, observations are allocated to cells in the two-way cross-classification, according to either a multinomial or Poisson sampling scheme. Suppose the sampling scheme is multinomial. Then at the second stage, for the (i,j) cell, each observation has probabilities $\lambda_{1(i)}$ or $\lambda_{2(j)}$ of losing its row or column identity, respectively, and thus becoming part of the corresponding row or column supplemental margins, respectively. Clearly an observation cannot become part of both supplemental margins, and we assume that there is a zero probability of an observation completely losing its identity. As a result an observation remains in its originally allocated cell, (i,j), with probability $1 - \lambda_{1(i)} - \lambda_{2(j)}$, so we must assume that $1 \geq \lambda_{1(i)} + \lambda_{2(j)}$ for all i and j. If the probability of allocation in the first stage to the (i,j) cell is π_{ij}, where $\sum_i \sum_j \pi_{ij} = 1$, then the probabilities associated with the cells at the end of the second stage are as illustrated for the 2 x 2 table in Table 1. Note that the sum of the probabilities over all eight cells is unity. For the presentation below we assume that $\lambda_{1(i)} > 0$ for $i = 1,2,\ldots,r$ and $\lambda_{2(j)} > 0$ for $j = 1,2,\ldots,c$. If some of the λ parameters are known to be zero, appropriate adjustments can be made.

<u>Table 1 goes about here</u>

If the original sampling scheme is a Poisson with expected value m_{ij} ($m_{ij} > 0$) for the (i,j) cell, then the parameters associated with the cells at the end of the second stage are similar to those illustrated for the 2 x 2 table in Table 1, with m_{ij}'s replacing π_{ij}'s.

3. MLE's for Poisson Sampling

Suppose we have an underlying Poisson sampling scheme generating observations at the unobservable first stage in the rc cells. Then at the end of the observable second stage we can think of there being an underlying Poisson scheme for the $rc + r + c = (r+1)(c+1) - 1$ cells in the $r \times c$ fully cross-classified table and the two sets of supplemental margins. The fact that there are actually $(r+1)(c+1) - 1$ cells with observed data will be quite important when we go to carry out goodness-of-fit tests of various models for the m's and the λ's.

The likelihood function for this second observable Poisson is proportional to

$$\exp(-\Sigma\,\Sigma\, m_{ij}) \prod_{i=1}^{r} \prod_{j=1}^{c} [(1 - \lambda_{1(i)} - \lambda_{2(j)}) m_{ij}]^{x_{ij}} \prod_{i=1}^{r} [\lambda_{1(i)} m_{i+}]^{R_i} \prod_{j=1}^{c} [\lambda_{2(j)} m_{+j}]^{C_j} \qquad (2)$$

where $m_{i+} = \sum_{j=1}^{c} m_{ij}$ and $m_{+j} = \sum_{i=1}^{r} m_{ij}$ for all i and j.

We note that (2) is a product of a function

$$f_1 = [\prod_{i=1}^{r} \prod_{j=1}^{c} (1 - \lambda_{1(i)} - \lambda_{2(j)})^{x_{ij}}][\prod_{i=1}^{r} \lambda_{1(i)}^{R_i}][\prod_{j=1}^{c} \lambda_{2(j)}^{C_j}] \qquad (3)$$

of the λ's, and a second function

$$f_2 = \exp(-\sum_{i=1}^{r} \sum_{j=1}^{c} m_{ij})[\prod_{i=1}^{r} \prod_{j=1}^{c} m_{ij}^{x_{ij}}][\prod_{i=1}^{r} m_{i+}^{R_i}][\prod_{j=1}^{c} m_{+j}^{C_j}] \qquad (4)$$

involving the m_{ij}'s. In maximizing the likelihood we can maximize each of these functions separately.

Our interest centers mainly on the m_{ij}'s because we would like to make inferences regarding the parameters in the first stage of the two-stage procedure, i.e. we would like to make inferences about the underlying cell parameters had there been no loss of information. We concern ourselves, however, with simplified models for the λ's since a reduction in the number of parameters affects degrees of freedom associated with goodness-of-fit tests, and because we are also interested in whether the supplemental margins "conform" to the observed margins (this information may be of value when we study related data with partially classified observations).

If any of the λ's are known to be zero, then we drop the appropriate terms from the likelihood function, (2). For example, if there are row supplemental marginal totals but no column ones, $\lambda_{2(j)} = 0$ for $j = 1,2,\ldots,c$, and the last part of the likelihood as given by (2), i.e. $\prod_j [\lambda_{2(j)} m_{+j}]^{C_j}$, is omitted. Dropping part of the likelihood is equivalent to setting the appropriate values of λ equal to zero in the various ML equations below.

For two-way tables we have two models of general interest for the m_{ij}'s: the unrestricted model which we label as $H_{(12)}$, and the model of independence of the variables corresponding to rows and to columns, i.e.

$$H_{(1,2)}: m_{ij} = m_{i+}m_{+j}/m_{++} \quad , \quad i = 1,\ldots,r, \; j = 1,\ldots,c. \quad (5)$$

We represent the unrestricted model in log-linear form as (see Birch [1963], Bishop and Fienberg [1969], or Fienberg [1970])

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad , \quad i = 1,\ldots,r, \; j = 1,\ldots,c, \quad (6)$$

where $\sum_{i=1}^{r} \alpha_i = 0$, $\sum_{j=1}^{c} \beta_j = 0$, $\sum_{i=1}^{r} \gamma_{ij} = \sum_{j=1}^{c} \gamma_{ij} = 0$. There is a one-to-one correspondence between the set of parameters $\{m_{ij}\}$ and the new set of parameters $\{\mu, \alpha_i, \beta_j, \gamma_{ij}\}$. The model of independence, $H_{(1,2)}$, can be represented in log-linear form by setting $\gamma_{ij} = 0$ for all i and j in (6).

When we take partial derivatives of the logarithm of the function f_2 in (3) with respect to μ, α_i, β_j, and γ_{ij}, we have the following:

$$\frac{\partial \log f_2}{\partial \mu} = -\sum_{i=1}^{r}\sum_{j=1}^{c} m_{ij} + \sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} + \sum_{i=1}^{r} R_i + \sum_{j=1}^{c} C_j \tag{7}$$

$$= -m_{++} + x_{++} + R_+ + C_+ ,$$

$$\frac{\partial \log f_2}{\partial \alpha_i} = -m_{i+} + x_{i+} + R_i + C_+\left(\frac{m_{i+}}{m_{++}}\right), \tag{8}$$

$$\frac{\partial \log f_2}{\partial \beta_j} = -m_{+j} + x_{+j} + R_+\left(\frac{m_{+j}}{m_{++}}\right) + C_j , \tag{9}$$

$$\frac{\partial \log f_2}{\partial \gamma_{ij}} = -m_{ij} + x_{ij} + R_i\left(\frac{m_{ij}}{m_{i+}}\right) + C_j\left(\frac{m_{ij}}{m_{+j}}\right) . \tag{10}$$

Under the unrestricted model, $H_{(12)}$, the maximum likelihood (ML) equations for the $\{m_{ij}\}$ are found by setting (7), (8), (9), and (10) equal to zero, i.e.

$$\hat{m}_{ij} = x_{ij} + R_i\left(\frac{\hat{m}_{ij}}{\hat{m}_{i+}}\right) + C_j\left(\frac{\hat{m}_{ij}}{\hat{m}_{+j}}\right), \quad i = 1,\ldots,r,\ j = 1,\ldots,c. \tag{11}$$

The left hand side of (11) is the expected count in the (i,j) cell after the unobservable first stage, while the right hand side is the observed count in the (i,j) cell plus a proportional allocation of the supplemental margins.

We cannot solve (11) for the $\{\hat{m}_{ij}\}$ in closed form, but the following iterative procedure will converge to the required maximum likelihood estimates (MLE's). As initial estimates of the $\{\hat{m}_{ij}\}$ we take

$$m_{ij}^{(0)} = \left(\frac{x_{ij}}{x_{++}}\right) N . \tag{12}$$

Note that if the data is fully cross-classified (i.e. $R_i = 0$ and $C_j = 0$, $i = 1,\ldots,r$ and $j = 1,\ldots,c$) the initial estimates are the usual unrestricted MLE's, $\{x_{ij}\}$. Then at the $(\nu+1)$ step (where $\nu \geq 1$) we take

$$m_{ij}^{(\nu+1)} = x_{ij} + R_i\left(\frac{m_{ij}^{(\nu)}}{m_{i+}^{(\nu)}}\right) + C_j\left(\frac{m_{ij}^{(\nu)}}{m_{+j}^{(\nu)}}\right) . \tag{13}$$

We continue iterating until the $m_{ij}^{(\nu)}$'s converge to a desired degree of accuracy. Actually, for any set of non-zero initial values, Np_{ij} could be used as initial estimates in place of (12), where $\sum_{ij} p_{ij} = 1$.

Under the model of independence, $H_{(1,2)}$, the ML equations are:

$$\begin{aligned} \hat{m}_{i+} &= x_{i+} + R_i + C_+\left(\frac{\hat{m}_{i+}}{\hat{m}_{++}}\right), \quad i = 1,2,\ldots,r, \\ \hat{m}_{+j} &= x_{+j} + R_+\left(\frac{\hat{m}_{+j}}{\hat{m}_{++}}\right) + C_j, \quad j = 1,2,\ldots,c. \end{aligned} \tag{14}$$

As with (11), the L.H.S. of (14) gives the expected numbers after the unobservable first stage, this time for the row and column margins, while the R.H.S. gives the corresponding observed counts plus the observed supplemental marginal counts plus proportional allocation of the other supplementary marginal total. Equations (14) are easily solved yielding

$$\hat{m}_{ij} = N\left(\frac{x_{i+} + R_i}{x_{++} + R_+}\right)\left(\frac{x_{+j} + C_j}{x_{++} + C_+}\right), \tag{15}$$

the intuitive estimates. When $\lambda_{1(i)} = \lambda_{2(j)} = 0$ for all i and j, $R_i = C_j = 0$ and (15) reduces to the usual cell probability estimates under independence.

Turning now to the λ's we consider four possible models:

(a) $S_{1(i),2(j)}$: the unrestricted model,

(b) $S_{1(i),2}$: $\lambda_{2(j)} = \lambda_2$, $j = 1,2,\ldots,c$,

(c) $S_{1,2(j)}$: $\lambda_{1(i)} = \lambda_1$, $i = 1,2,\ldots,r$,

(d) $S_{1,2}$: $\lambda_{1(i)} = \lambda_1$, $i = 1,2,\ldots,r$

$\lambda_{2(j)} = \lambda_2$, $j = 1,2,\ldots,c$.

We refer to $S_{1(i),2}$ ($S_{1,2(j)}$) as the model of "marginal conformity" of the column (row) supplemental margins. $S_{1,2}$ is simply a combination of the two parameter reductions, (b) and (c).

Surprisingly, MLE's under the unrestricted model $S_{1(i),2(j)}$ are the most complicated to compute. In this case the ML equations based on (3) are

$$\frac{R_i}{\hat{\lambda}_{1(i)}} = \sum_{j=1}^{c} \frac{x_{ij}}{1 - \hat{\lambda}_{1(i)} - \hat{\lambda}_{2(j)}} \quad , \quad i = 1,2,\ldots,r, \tag{16}$$

and

$$\frac{C_j}{\hat{\lambda}_{2(j)}} = \sum_{i=1}^{r} \frac{x_{ij}}{1 - \hat{\lambda}_{1(i)} - \hat{\lambda}_{2(j)}} \quad , \quad j = 1,2,\ldots,c, \tag{17}$$

and we require iteration for a solution. As initial estimates we take

$$\lambda_{1(i)}^{(0)} = R_+/N \ , \ \lambda_{2(j)}^{(0)} = C_+/N \tag{18}$$

and, if $R_i \leq x_{i+}$ for all i and $C_j \leq x_{+j}$ for all j, at the $(\nu+1)$ stage $(\nu \geq 1)$ we set

$$\lambda_{1(i)}^{(\nu+1)} = R_i / \sum_{j=1}^{c} \left(\frac{x_{ij}}{1 - \lambda_{1(i)}^{(\nu)} - \lambda_{2(j)}^{(\nu)}} \right) \tag{19}$$

and

$$\lambda_{2(j)}^{(\nu+1)} = C_j / \sum_{i=1}^{r} \left(\frac{x_{ij}}{1 - \lambda_{1(i)}^{(\nu)} - \lambda_{2(j)}^{(\nu)}} \right) \tag{20}$$

continuing until the values at successive steps are sufficiently close. If for some k, $R_k > x_{k+}$, then at some stage $\lambda_{1(k)}^{(\nu)}$ will fall outside the range [0,1] and we must replace (20) for that value of i by

$$\lambda_{1(k)}^{(\nu+1)} = 1 - \lambda_{2(\ell)}^{(\nu)} - \frac{\lambda_{1(k)}^{(\nu)}}{R_k} \left[\sum_{j=1}^{c} \frac{x_{kj}}{1 - \lambda_{1(k)}^{(\nu)} - \lambda_{2(j)}^{(\nu)}} \right] (1 - \lambda_{1(k)}^{(\nu)} - \lambda_{(\ell)}^{(\nu)}) \tag{21}$$

where ℓ is a subscript chosen at each step such that $\lambda_{2(\ell)}^{(\nu)} \geq \lambda_{2(j)}^{(\nu)}$ for all j. Similarly if $C_h > x_{+h}$ for some h, we replace the corresponding equation in (20) by one analogous to (21).

For the model $S_{1(i),2}$, marginal conformity for the column supplemental margins, the ML equations reduce to

$$\frac{R_i}{\hat{\lambda}_{1(i)}} = \frac{x_{i+}}{1 - \hat{\lambda}_{1(i)} - \hat{\lambda}_2} \tag{22}$$

$$\frac{C_+}{\hat{\lambda}_2} = \sum_{i=1}^{r} \frac{x_{i+}}{1 - \hat{\lambda}_{1(i)} - \hat{\lambda}_2} \ . \tag{23}$$

These equations have the explicit solution

$$\hat{\lambda}_{1(i)} = \left(\frac{x_{++} + R_+}{N}\right)\left(\frac{R_i}{x_{i+} + R_i}\right) \ , \quad i = 1,2,\ldots,r, \tag{24}$$

and

$$\hat{\lambda}_2 = C_+/N \ . \tag{25}$$

For $S_{1,2(j)}$ the MLE's are similar to (24) and (25), with the roles of rows and columns reversed, and for $S_{1,2}$ the MLE's are $\hat{\lambda}_1 = R_+/N$ and $\hat{\lambda}_2 = C_+/N$, the intuitive estimates.

Once we have MLE's for the $\{m_{ij}\}$ and the λ's, we can compute expected values for the (r+1)(c+1) - 1 cells with observed counts. If $\{\theta_{ij}\}$ are the expected counts for the cross-classified cells, $\{\rho_i\}$ the expected values for the row supplemental margins, and $\{\sigma_j\}$ the expected values for the column supplemental margins, then under $S_{1(i),2(j)}$

$$\hat{\theta}_{ij} = N(1 - \hat{\lambda}_{1(i)} - \hat{\lambda}_{2(j)})\hat{m}_{ij} \tag{26}$$

$$\hat{\rho}_i = N\,\hat{\lambda}_{1(i)}\hat{m}_{i+} \tag{27}$$

$$\hat{\sigma}_j = N\,\hat{\lambda}_{2(j)}\hat{m}_{+j} \ . \tag{28}$$

When we are considering $S_{1,2(j)}$, $S_{1(i),2}$ or $S_{1,2}$ we replace the appropriate λ's accordingly.

There are eight possible models for the data once we combine the models for the $\{m_{ij}\}$ and for the λ's. Since $S_{1(i),2}$ is similar in form to $S_{1,2(j)}$, we need only consider six of the eight models. Clearly, for $[H_{(12)} + S_{1(i),2(j)}]$

we are fitting all the degrees of freedom and $\hat{\theta}_{ij} = x_{ij}$, $\hat{\rho}_i = R_i$ and $\hat{\sigma}_j = C_j$. Because there are no explicit formulae for the $\hat{\lambda}$'s under $S_{1(i),2(j)}$ we have no explicit form for the estimated expected values under $[H_{(12)} + S_{1(i),2(j)}]$, and, although the goodness-of-fit test statistics for this model have $rc + r + c - 1 - (r-1) - (c-1) - r - c = (r-1)(c-1)$ d.f. as we might have expected, these statistics are not the usual ones we use to test independence. Similarly, since there are no explicit values for the m_{ij}'s under $H_{(12)}$, we have no explicit form for the estimated expected values under $[H_{(12)} + S_{1,2(j)}]$ or under $[H_{(12)} + S_{1,2}]$. The d.f. for these two models are $(r-1)$ and $(r+c-2)$, respectively. Finally, under $[H_{(12)} + S_{1,2(j)}]$ we have

$$\hat{\theta}_{ij} = \left(\frac{x_{i+} + R_i}{x_{++} + R_+}\right) x_{+j} \tag{29}$$

$$\hat{\rho}_i = R_+\left(\frac{x_{i+} + R_i}{x_{++} + R_+}\right) \tag{30}$$

$$\hat{\sigma}_j = C_j \tag{31}$$

with $rc + r + c - 1 - (r-1) - (c-1) - c - 1 = (r-1)c$ d.f., while under $[H_{(12)} + S_{1,2}]$ we have

$$\hat{\theta}_{ij} = x_{++}\left(\frac{x_{i+} + R_i}{x_{++} + R_+}\right)\left(\frac{x_{+j} + C_j}{x_{++} + C_+}\right) \tag{32}$$

$$\hat{\rho}_i = R_+\left(\frac{x_{i+} + R_i}{x_{++} + R_+}\right) \tag{33}$$

$$\hat{\sigma}_j = C_+\left(\frac{x_{+j} + C_j}{x_{++} + C_+}\right) \tag{34}$$

with $rc-1$ d.f.

4. MLE's for Multinomial Sampling

Suppose we have an underlying multinomial sampling scheme allocating observations at the unobservable first stage to the rc cells. Then at the end of the observable second stage we can think of there being an underlying multinomial scheme for the (r+1)(c+1) - 1 cells in the r x c fully cross-classified table and the two sets of supplemental margins. The likelihood function for this second observable multinomial is proportional to

$$\prod_{i=1}^{r} \prod_{j=1}^{c} [(1 - \lambda_{1(i)} - \lambda_{2(j)})\pi_{ij}]^{x_{ij}} \cdot \prod_{i=1}^{r} [\lambda_{1(i)}\pi_{i+}]^{R_i} \cdot \prod_{j=1}^{c} [\lambda_{2(j)}\pi_{+j}]^{C_j}, \tag{35}$$

which can also be written as a product of two functions, one involving only the λ's, which is given by expression (3), and a second involving only the π_{ij}'s:

$$[\prod_{i=1}^{r} \prod_{j=1}^{c} \pi_{ij}^{x_{ij}}][\prod_{i=1}^{r} \pi_{i+}^{R_i}][\prod_{j=1}^{c} \pi_{+j}^{C_j}]. \tag{36}$$

Since the part of the likelihood function involving the λ's is the same as in the preceding section, the MLE's derived there are also appropriate for the multinomial sampling scheme. From the ML equations (11) and (14) we see that for the Poisson sampling scheme

$$\hat{m}_{++} = x_{++} + R_+ + C_+ = N. \tag{37}$$

Using an argument similar to that given by Birch [1963], we can show that the MLE's of $\{\pi_{ij}\}$ under the multinomial sampling model are

$$\hat{\pi}_{ij} = \hat{m}_{ij}/N \tag{38}$$

where $\{\hat{m}_{ij}\}$ are the corresponding MLE's under the Poisson sampling model. For example, under the model of independence

$$H_{(1,2)}: \pi_{ij} = \pi_{i+}\pi_{+j} \quad , \quad i = 1,\ldots,r, \ j = 1,\ldots,c, \tag{39}$$

the MLE's of $\{\pi_{ij}\}$ are

$$\hat{\pi}_{ij} = \left(\frac{x_{i+} + R_i}{x_{++} + R_+}\right)\left(\frac{x_{+j} + C_j}{x_{++} + C_+}\right). \tag{40}$$

Keeping in mind the correspondence given by (38), we can use the expected cell values derived in the preceding section for both multinomial and Poisson situations.

In Section 1 we mentioned the approach of Blumenthal [1968] and Hocking and Oxspring [1971]. They treat the completely classified data as being multinomially distributed, and the partially classified data as consisting of one or more multinomial observations whose parameters are directly related to those of the multinomial for the completely classified data. For our problem this approach would yield three multinomials: one for the completely cross-classified data, one for the row supplemental margins, and one for the column supplemental margins. The combined likelihood is given by expression (36), and the MLE's for the underlying cell probabilities are given by (38) or by (40). The goodness-of-fit statistics described in the next section are still appropriate in this case.

5. Goodness-of-Fit Tests

We can carry out goodness-of-fit tests for the models described above using either the Pearson or the likelihood ratio statistics:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - \hat{\theta}_{ij})^2}{\hat{\theta}_{ij}} + \sum_{i=1}^{r} \frac{(R_i - \hat{\rho}_i)^2}{\hat{\rho}_i} + \sum_{j=1}^{c} \frac{(C_j - \hat{\sigma}_j)^2}{\hat{\sigma}_j} \tag{41}$$

or

$$G^2 = 2\left[\sum_{i=1}^{r} \sum_{j=1}^{c} x_{ij} \log \frac{x_{ij}}{\hat{\theta}_{ij}} + \sum_{i=1}^{r} R_i \log \frac{R_i}{\hat{\rho}_i} + \sum_{j=1}^{c} C_j \log \frac{C_j}{\hat{\sigma}_j}\right], \tag{42}$$

each having an asymptotic χ^2 distribution with degrees of freedom as indicated above, under the appropriate null model.

Note that under $[H_{(1,2)} + S_{1,2(j)}]$ the third term on the R.H.S. of (41) and (42) is zero, and the values of the $\{\hat{\sigma}_j\}$ do not effect the $\{\hat{\theta}_{ij}\}$. The resulting test statistic in this case is the same as the one we would have derived had we considered the case of one supplemental margin rather than two.

The models considered above have a special hierarchical structure which suggests a way to partition various goodness-of-fit statistics, i.e.

$$\begin{array}{ccccc} [H_{(12)} + S_{1(i),2(j)}] & \supset & [H_{(12)} + S_{1,2(j)}] & \supset & [H_{(12)} + S_{1,2}] \\ \cup & & \cup & & \cup \\ [H_{(1,2)} + S_{1(i),2(j)}] & \supset & [H_{(1,2)} + S_{1,2(j)}] & \supset & [H_{(1,2)} + S_{1,2}] \end{array} \tag{43}$$

where we use the notation $A \supset B$ to mean that model B is a special case of model A. Thus we can consider the following exact partitions of the likelihood-ratio statistic for the simplest model (the one with the smallest number of parameters), $[H_{(1,2)} + S_{1,2}]$, in terms of various conditional and unconditional likelihood-ratio test statistics:

$$G^2[H_{(1,2)} + S_{1,2}] = G^2[H_{(12)} + S_{1,2}] + G^2[H_{(1,2)} + S_{1,2}|H_{(12)} + S_{1,2}]$$

$$= G^2[H_{(12)} + S_{1,2(j)}] + G^2[H_{(12)} + S_{1,2}|H_{(12)} + S_{1,2(j)}]$$

$$+ G^2[H_{(1,2)} + S_{1,2}|H_{(12)} + S_{1,2}] \tag{44}$$

or

$$G^2[H_{(1,2)} + S_{1,2}] = G^2[H_{(1,2)} + S_{1,2(j)}] + G^2[H_{(1,2)} + S_{1,2}|H_{(1,2)} + S_{1,2(j)}]$$

$$= G^2[H_{(1,2)}+S_{1(i),2(j)}] + G^2[H_{(1,2)}+S_{1,2(j)}|H_{(1,2)}+S_{1(i),2(j)}]$$

$$+ G^2[H_{(1,2)} + S_{1,2}|H_{(1,2)} + S_{1,2(j)}] \tag{45}$$

or

$$G^2[H_{(1,2)} + S_{1,2}] = G^2[H_{(12)} + S_{1,2(j)}] + G^2[H_{(1,2)} + S_{1,2(j)}|H_{(12)} + S_{1,2(j)}]$$

$$+ G^2[H_{(1,2)} + S_{1,2}|H_{(1,2)} + S_{1,2(j)}] . \tag{46}$$

We will make use of such partitions in an example below.

6. Ignoring the Partially Classified Data

One way to handle data of the form described in this paper is simply to ignore the partially classified data and analyse only the cross-classified counts $\{x_{ij}\}$. Let us consider the multinomial sampling model. If model $S_{1,2}$, the marginal conformity of the supplemental margins, is not true then the usual estimates for the $\{\pi_{ij}\}$ based on ignoring the partially classified data are statistically inconsistent (i.e. as the sample N tends to ∞, they do not converge to the true values of the $\{\pi_{ij}\}$). This is immediately clear if we note that the expected cell values for the $\{x_{ij}\}$ under $S_{1(i),2(j)}$ or $S_{1,2(j)}$ involve different λ values for different cells. Thus, ignoring the supplemental margins can lead to incorrect results. If, on the other hand, $S_{1,2}$ is true then the usual estimates under $H_{(12)}$ or $H_{(1,2)}$ are consistent; however, the estimates proposed here are more efficient, in the sense that they have smaller asymptotic variances (see Chen [1972]). For example, in the 2 x 2 table the asymptotic variance of $\hat{\pi}_{11}$ under $[H_{(12)} + S_{1,2}]$ is equal to

$$\frac{\pi_{11}(1-\pi_{11})}{N(1-\lambda_1-\lambda_2)} - \frac{\lambda_1\pi_{11}^2\pi_{2+}}{N_\beta(1-\lambda_1-\lambda_2)\pi_{1+}} - \frac{\lambda_2\pi_{11}^2\pi_{+2}}{N_\beta(1-\lambda_1-\lambda_2)\pi_{+1}}$$
$$- \frac{\lambda_1\lambda_2}{N_\beta(1-\lambda_1-\lambda_2)^2}\left(\frac{\pi_{11}\pi_{12}\pi_{21}\pi_{22}}{\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2}}\right)\left[\pi_{11}(1-\pi_{11})\left(\frac{1}{\pi_{11}}+\frac{1}{\pi_{12}}+\frac{1}{\pi_{21}}+\frac{1}{\pi_{22}}\right)-1\right]. \tag{47}$$

where

$$N_\beta = N\left[1+\frac{\lambda_1\lambda_2}{(1-\lambda_1-\lambda_2)}\,\frac{\pi_{11}\pi_{12}\pi_{21}\pi_{22}}{\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2}}\left(\frac{1}{\pi_{11}}+\frac{1}{\pi_{12}}+\frac{1}{\pi_{21}}+\frac{1}{\pi_{22}}\right)\right]. \tag{48}$$

Since the last three terms are negative, the asymptotic variance of $\hat{\pi}_{11}$ is smaller than

$$\frac{\pi_{11}(1-\pi_{11})}{N(1-\lambda_1-\lambda_2)} \sim \frac{\pi_{11}(1-\pi_{11})}{x_{++}}. \tag{49}$$

But the R.H.S. of (49) is the variance of the usual estimate of π_{11}, x_{ij}/x_{++}, based on ignoring the supplemental margins.

7. An Example

Reinfurt [1970] considers data on 456 premature live births (i.e. infant's birth weight was less than or equal to 2000 grams), which we give in Table 2. The five-minute Apgar index is a composite of heart rate, respiratory effort, reflex irritability, muscle tone and color as observed five minutes after birth, each component receiving a score of 0, 1 or 2. Thus the index range is 0-10 with lower scores indicating healthier infants. A serum bilirium count exceeding 1.0 mg. per 100 ml. is indicative of a malfunctioning kidney, usually accompanied by considerable jaundice, a condition which, if allowed to persist, can lead to damage of the central nervous system. Of the 456 observations, 153 are partially classified according to five-minute Apgar score and 24 are partially classified according to serum bilirium level.

We begin our analysis by exploring the structure of the λ parameters. Under $[H_{(12)} + S_{1(i),2}]$ we compute our estimates of the $\{m_{ij}\}$ using (12) and (13). After four cycles of the iteration we get, to two-decimal accuracy,

$$\hat{m}_{11} = 85.73 \ , \ \hat{m}_{12} = 95.30 \ , \ \hat{m}_{21} = 134.98 \ , \ \hat{m}_{22} = 140.90. \tag{50}$$

The MLE's of $\lambda_{1(1)}$, $\lambda_{1(2)}$ and λ_2 can be written directly using (24) and (25):

$$\hat{\lambda}_{1(1)} = 0.060 \ , \ \hat{\lambda}_{1(2)} = 0.047 \ , \ \hat{\lambda}_2 = 0.334. \tag{51}$$

Combining (50) and (51) we get the estimated expected values:

$$\hat{\theta}_{11} = 53.25 \ , \ \hat{\theta}_{12} = 57.68 \ , \ \hat{\theta}_{21} = 82.10 \ , \ \hat{\theta}_{22} = 86.71 \ ;$$
$$\hat{\rho}_1 = 11.01 \ , \ \hat{\rho}_2 = 12.90 \ ; \ \hat{\sigma}_1 = 73.70 \ , \ \hat{\sigma}_2 = 78.64 \ .$$

Note that $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are not exactly equal to C_1 and C_2 as we might have expected. They are quite close, however, and the deviations contribute little to the goodness-of-fit statistics. The goodness-of-fit statistic values for this model are $X^2 = 75.06$ and $G^2 = 77.79$, with 1 d.f. Clearly this model does not describe the data.

Turning to $[H_{(12)} + S_{1,2(j)}]$, we use the same estimates of the $\{m_{ij}\}$ while the MLE's for the λ's are now

$$\hat{\lambda}_1 = 0.053 \ , \ \hat{\lambda}_{2(1)} = 0.530 \ , \ \hat{\lambda}_{2(2)} = 0.153 \ .$$

The corresponding expected cell counts are

$$\hat{\theta}_{11} = 35.70 \ , \ \hat{\theta}_{12} = 75.78 \ , \ \hat{\theta}_{21} = 56.28 \ , \ \hat{\theta}_{22} = 111.25 \ ;$$
$$\hat{\rho}_1 = 9.53 \ , \ \hat{\rho}_2 = 14.47 \ ; \ \hat{\sigma}_1 = 116.99 \ , \ \hat{\sigma}_2 = 36.00$$

and $X^2 = 0.41$ and $G^2 = 0.41$, with 1 d.f. Thus model $S_{1,2(j)}$ seems highly appropriate for the data, and since $S_{1(i),2}$ seemed so inappropriate we do not consider $S_{1,2}$.

Having decided on a model for the λ's, we can explore the relationship between Apgar score and serum bilirium level. Under $H_{(1,2)}$, independence, $\hat{m}_{1+} = (N - \hat{m}_{2+}) = 182.4$ and $\hat{m}_{+1} = (N - \hat{m}_{+2}) = 218.88$; so the expected cell counts under $[H_{(1,2)} + S_{1,2(j)}]$ are

$$\hat{\theta}_{11} = 36.86 \ , \ \hat{\theta}_{12} = 74.92 \ , \ \hat{\theta}_{21} = 55.14 \ , \ \hat{\theta}_{22} = 112.09 \ ;$$
$$\hat{\rho}_1 = 9.62 \ , \ \hat{\rho}_2 = 14.39 \ ; \ \hat{\sigma}_1 = 117.00 \ , \ \hat{\sigma}_2 = 36.00 \ .$$

The goodness-of-fit test statistic values are $X^2 = 0.49$ and $G^2 = 0.49$, each with 2 d.f. Moreover, the conditional statistic $G^2[H_{(1,2)}+S_{1,2(j)} | H_{(12)}+S_{1,2(j)}] = 0.08$, with 1 d.f.

We conclude that the Apgar score and the serum bilirium level are unrelated in these premature infants, and that those infants without a serum bilirium reading are more likely to have a reading of 1.0 or below. Thus the 2 x 2 table of fully cross-classified data would appear to be quite deceptive, and the actual probability of a low Apgar score <u>and</u> low serum bilirium reading is not very much smaller than the probability of a high score <u>and</u> high reading.

References

Birch, M.W. [1963]. Maximum likelihood in three-way contingency tables. J. Roy. Statist. Soc. Ser. B 25, 220-233.

Bishop, Y.M.M. [1969]. Full contingency tables, logits, and split contingency tables. Biometrics 25, 383-389.

Bishop, Y.M.M. and Fienberg, S.E. [1969]. Incomplete two-dimensional contingency tables. Biometrics 25, 119-128.

Blumenthal, S. [1968]. Multinomial sampling with partially categorized data. J. Amer. Statist. Assoc. 63, 542-551.

Chen, T. [1972]. Mixed-up frequencies and missing data in contingency tables. Unpublished Ph.D. dissertation, Department of Statistics, University of Chicago.

Fienberg, S.E. [1970]. The analysis of multidimensional contingency tables. Ecology 51, 419-433.

Hocking, R.R. and Oxspring, H.H. [1971]. Maximum likelihood estimation with incomplete multinomial data. J. Amer. Statist. Assoc. 66, 65-70.

Koch, G.G. and Reinfurt, D.W. [1970]. The analysis of complex contingency table data from general experimental designs and sample surveys. Proceedings of the Sixteenth Conference on the Design of Experiments in Army Research, Development and Testing, Fort Lee, Virginia.

Koch, G.G., Imrey, P.B., and Reinfurt, D.W. [1972]. Linear model analysis of categorical data with incomplete response vectors. Biometrics 28, 663-692.

Reinfurt, D.W. [1970]. The analysis of categorical data with supplemented margins including applications to mixed models. Unpublished Ph.D. dissertation, Department of Biostatistics, University of North Carolina.

Table 1

Underlying Probabilities for a 2 x 2 Table
with Two Sets of Partially Classified Margins

	Fully Classified Table 1	2	Row Supplemental Margin
1	$(1-\lambda_{1(1)}-\lambda_{2(1)})\pi_{11}$	$(1-\lambda_{1(1)}-\lambda_{2(2)})\pi_{12}$	$\lambda_{1(1)}\pi_{1+}$
2	$(1-\lambda_{1(2)}-\lambda_{2(1)})\pi_{21}$	$(1-\lambda_{1(2)}-\lambda_{2(2)})\pi_{22}$	$\lambda_{1(2)}\pi_{2+}$
Column Supplemental Margin	$\lambda_{2(1)}\pi_{+1}$	$\lambda_{2(2)}\pi_{+2}$	

Table 2

Data of Premature Infants Cross-classified According to Apgar Index and Serum Bilirubin Level with Supplemental Margins (Reinfurt [1970])

Serum Bilirubin Reading	5-Minute Apgar Score 0-6	7-10	Subtotal	Supplementation on Serum Bilirubin Reading	Total
0-1.0	35	75	110	11	121
$\geq$ 1.1	57	112	169	13	182
Subtotal	92	187	279	24	303
Supplementation on Apgar Score	117	36	153		
Total	209	223	432		456